M.Com (Ist Sem) Examination, 2013

Statistical Analysis

[MODEL ANSWER]

1. Answer the following questions:

(i) The movements which exhibit persistant growth or decline in a time series is known as Linear Trend.

For a given time series data, to obtain a linear trend, the values of 'a' and 'b' are obtained by the normal equations

$$\Sigma y = na + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

where n = number of pairs of data given.

a is the intercept of the line on y-axis and

b is the slope of the line.

The mild or violent movements that are random in nature and the variations are produced by occasional forces are called Non-Linear Trend. The trend has no definite pattern and confined to very short period of time.

Forms of non-linear equations are:

BSir

$Y = a + bx + cx^2$ (parabola of second degree)

$Y = a + bx + cx^2 + dx^3$ (parabola of third degree)

(ii)(iii) **Calculation of co-efficient of correlation :-**

Given $\bar{X} = 6$, $\bar{Y} = 8$

X :	6	2	10	a	8
Y :	9	11	b	8	7

For x series, $\bar{X} = \frac{6+2+10+a+8}{5}$

$5\bar{X} = 26 + a$

$\therefore 5(6) = 26 + a$

$\therefore \boxed{a = 4}$

For y series, $\bar{Y} = \frac{9+11+b+8+7}{5}$

$5\bar{Y} = 35 + b$

$\therefore 5(8) = 35 + b$

$\therefore \boxed{b = 5}$

X	$x = X - \bar{X}$	x^2	Y	$y = Y - \bar{Y}$	y^2	xy
6	0	0	9	1	1	0
2	-4	16	11	3	9	-12
10	4	16	5	-3	9	-12
4	2	4	8	0	0	0
8	2	4	7	-1	1	-2
	0	40		0	20	-26

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{-26}{\sqrt{40 \times 20}} = \frac{-26}{28.28}$$

$$\therefore \boxed{r = -0.92}$$

(iii) Regression Analysis :

The term regression analysis refers to the methods by which estimates are made or predicted for dependent variable by using the independent variables having functional relationship.

Uses of regression analysis in business problems:

a) Indicate relationship between two variables.

ex- Most likely price in Delhi of a particular commodity, corresponding to a certain price of same commodity at Kanpur.

b) To measure a change in one variable with a unit change in other variable.

ex- Estimate the production cost per unit of commodity with increase in labour cost of Rs 10 per person.

Beju

(iv) **Bernoulli's theorem of Probability.**

$$(q+p)^n = q^n + {}^nC_1 q^{n-1}p + {}^nC_2 q^{n-2}p^2 + \ldots\ldots + \ldots\ldots + {}^nC_r q^{n-r}p^r$$

where, n = number of independent trials.

p = Probability of success in a single trial.

q = Probability of failure in a single trail.

r = Number of successes

$n-r$ = Number of failures.

(v) The values can be obtained by a Nine square table

	A	α	Total
B	(AB) 200	(αB) 150	(B) 350
β	(Aβ) 25	($\alpha\beta$) 125	(β) 150
Total	(A) 225	(α) 275	N 500

$\therefore$ (A) = 225, (α) = 275, (Aβ) = 25

(vi) No. of playing cards in a ordinary pack = 52

Probability of drawing a black card = $\frac{26}{52}$

Probability of drawing a king = $\frac{4}{52}$

$\therefore$ Probability of drawing a black card or a king = $\frac{26}{52} + \frac{4}{52} - \frac{2}{52} = \frac{7}{13}$.

(vii) Stratified Random Sampling:

In the process of sampling, when a population is heterogeneous, it is divided into groups on basis of some common factor, so that each group (called stratum) is homogeneous. The strata are non-overlapping and they comprise the whole population. Such sample is called random when each stratum is done randomly. Thus, samples obtained in the above process is called Stratified Random Sampling.

(viii) Hypothesis: The ratio of boys and girls is 1:1 in new born children.

N = Total new born children = 800

∴ As per hypothesis the expected no. of boys and girls = 400 each.

Computation of χ^2

	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Boys	500	400	+100	10,000	25
Girls	300	400	−100	10,000	25
	800	800			$\chi^2 = 50$

d.f. = (N−1) = 2−1 = 1

The table value of χ^2 at 5% level of significance with 1 d.f. is 3.841

$\chi^2_{cal} > \chi^2_{tab.}$

∴ The hypothesis is not true.

(ix) **Coefficient of Calligration:**

For the techniques of Association of Attributes and Testing a formula developed by Prof. Yule is known as coefficient of calligration and is given by

$$Y_{AB} = \frac{1 - \sqrt{\frac{(AB)(\alpha\beta)}{(A\beta)(\alpha B)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}$$

(x) **t-test**: one of the most important test of Significance in case of small samples is t-test which is based on t-distribution, discovered by William Gosset.

t-test is used to test the significance between difference of sample mean and population mean or to test the null hypothesis $\mu = \mu_0$

$$\boxed{t = \frac{\bar{x} - \mu}{S}}$$

where $\bar{x}$ = sample mean = $\frac{\Sigma x}{n}$

μ = Population mean

S = standard deviation of samples = $\sqrt{\frac{\Sigma d^2}{n-1}}$

BSJ

Table value of t is calculated with

$$d.f. = (n-1)$$

$\therefore t_{cal} > t_{tab}$, difference is significant and Ho, null hypothesis is rejected.

f-test: It is also called variance ratio test, developed by G.W. Snedecore, to test the equality of two variances on the basis of two independent samples from two normal populations.

$$F = \frac{S_1^2}{S_2^2}$$

where, $S_1^2 = \frac{\Sigma (x_1 - \bar{x}_1)^2}{n_1 - 1}$, $S_2^2 = \frac{\Sigma (x_2 - \bar{x}_2)^2}{n_2 - 1}$

Table value of f is calculated with (v_1, v_2) d.f.

where $v_1 = n_1 - 1$, $v_2 = n_2 - 1$

$\therefore f_{cal} > f_{tab}$, then the ratio is significant.

and $f_{cal} < f_{tab}$, then the ratio is not significant which means the two samples are taken from the ~~same~~ population having same variances.

2 Solution:

Year	Price (Y)	Year−2008 (x)	x^2	x^3	x^4	xy	x^2y
2005	13	−3	9	−27	81	−39	117
2006	13	−2	4	−8	16	−26	52
2007	22	−1	1	−1	1	−22	22
2008	21	0	0	0	0	0	0
2009	54	1	1	1	1	54	54
2010	60	2	4	8	16	120	240
2011	83	3	9	27	81	249	747
n=7	266	$\Sigma x=0$	28	0	196	336	1,232

Let $y = a+bx+cx^2$

Then normal equations are:

$$\Sigma y = Na+b\Sigma x+c\Sigma x^2$$
$$\Sigma xy = a\Sigma x+b\Sigma x^2+c\Sigma x^3$$
$$\Sigma x^2y = a\Sigma x^2+b\Sigma x^3+c\Sigma x^4$$

Since $\Sigma x=0=\Sigma x^3$, the reduced equations are:

$$\Sigma y = Na+c\Sigma x^2 \quad \text{—(i)}$$
$$\Sigma xy = b\Sigma x^2 \quad \text{—(ii)}$$
$$\Sigma x^2y = a\Sigma x^2+c\Sigma x^4 \quad \text{—(iii)}$$

Substituting the values

$$266 = 7a+28c \quad \text{—(1)}$$
$$336 = 28b, \boxed{b=12} \quad \text{—(2)}$$

$$1,232 = 28a + 196c \quad \text{——(3)}$$

Multiplying eq (1) by 4

$$1,064 = 28a + 112c \quad \text{——(4)}$$

Subtracting eq (4) from eq (3)

$$\begin{array}{r} 1,232 = 28a + 196c \\ 1,064 = 28a + 112c \\ \hline 168 = \quad 84c \end{array}$$

$$\therefore \quad c = \frac{168}{84} = 2, \quad \boxed{c = 2}$$

Substituting the value of c in eq(1)

$$266 = 7a + 28(2)$$

$$a = \frac{210}{70} = 30, \quad \boxed{a = 30}$$

Substituting the values of a, b, c in the parabola,

$$Y = 30 + 12x + 2x^2$$

The Trend values are obtained as follows:

Year	x	$T = 30 + 12x + 2x^2$
2005	-3	$30 + 12(-3) + 2 \times 9 = 12$
2006	-2	$30 + 12(-2) + 2 \times 4 = 14$
2007	-1	$30 + 12(-1) + 2 \times 1 = 20$

Year	x	Trend
2008	0	$30+(12\times0)+(2\times0)=30$
2009	1	$30+(12\times1)+(2\times1)=44$
2010	2	$30+(12\times2)+(2\times4)=62$
2011	3	$30+(12\times3)+(2\times9)=84$

Trend value for the year 2015 :

$$Y_{2015} = 30+(12\times7)+(2\times7^2)$$
$$= 30+84+98$$
$$= 212$$

(where $x = 2015-2008$)
ie.. $x=7$

---

3 Solution :

$$\text{Average income } (x) = \frac{\text{Total income}}{\text{No. of persons}}$$

$$\text{Average expenditure } (Y) = \frac{\text{Total expenditure}}{\text{No. of persons}} \quad \text{(Given)}$$

x	dx	dx^2	Y	dy	dy^2	$dxdy$
80	30	900	12	-6	36	-180
70	20	400	16	-2	4	-40
60	10	100	18	0	0	0
40	-10	100	19	1	1	-10
30	-20	400	21	3	9	-60
20	-30	900	22	4	16	-120
$N=6$	$\Sigma dx=0$	Σdx^2 $=2800$		$\Sigma dy=0$	Σdy^2 $=66$	$\Sigma dxdy$ $=-410$

$[dx = x - \bar{x}]$ $\quad$ $[dy = y - \bar{Y}]$

$\bar{x} = \frac{300}{6} = \underline{50}$ $\quad$ $\bar{Y} = \frac{108}{6} = \underline{18}$

$$r = \frac{\Sigma dxdy - \frac{(\Sigma dx)(\Sigma dy)}{N}}{\sqrt{\left[\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}\right]\left[\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}\right]}}$$

$$= \frac{-410 - 0}{\sqrt{[2800]\,[66]}} = \frac{-410}{\sqrt{184800}} = \frac{-410}{429.88}$$

$\therefore \quad r = -0.95$

High degree negetive correlation.

4 Seven dice are thrown simultaneously,

$\therefore n = 7$

Probability of getting "1" on a dice = $1/6 = p$

probability of not getting "1" on a dice = $5/6 = q$

(a) Probability of getting "1" 4 times = ${}^nC_r (p)^r (q)^{n-r}$

$${}^7C_4 (p)^4 (q)^3$$

$$= {}^7C_4 (1/6)^4 (5/6)^3$$

$$= 35 \times \frac{1}{1296} \times \frac{125}{216} = \frac{4375}{279936} = \boxed{0.015}$$

(b) Probability of getting "1" at least 3 times =

$${}^7C_3 (1/6)^3 (5/6)^4 + {}^7C_4 (1/6)^4 (5/6)^3 + {}^7C_5 (1/6)^5 (5/6)^2 + {}^7C_6 (1/6)^6 (5/6)^1 + {}^7C_7 (1/6)^7 (5/6)^0$$

$$= 35 \times \frac{625}{279936} + 35 \times \frac{125}{279936} + 21 \times \frac{25}{279936} + 7 \times \frac{5}{279936} + \frac{1}{279936}$$

$$= \frac{21875}{279936} + \frac{4375}{279936} + \frac{525}{279936} + \frac{35}{279936} + \frac{1}{279936}$$

$$= \frac{26810+1}{279936} = \boxed{0.095}$$

(c) Probability of getting "3" on 4 dice at least =

$${}^7C_4 (1/6)^4 (5/6)^3 + {}^7C_5 (1/6)^5 (5/6)^2 + {}^7C_6 (1/6)^6 (5/6)^1 + {}^7C_7 (1/6)^7 (5/6)^0$$

$$\frac{4375}{279936} + \frac{525}{279936} + \frac{35}{279936} + \frac{1}{279936} = \frac{4936}{279936} = \boxed{0.017}$$

5 Let, H_0 = "The education level depends upon sex" be true.

∴ Expected Frequency table :

Sex	H.S (A_1)	UG (A_2)	PG (A_3)	Total
Male (B_1)	12	18	30	60 (B_1)
Female (B_2)	30	12	18	60 (B_2)
	42 (A_1)	30 (A_2)	48 (A_3)	120 (N)

Association	Calculation	Exp. freq.	Obs. freq.
(A_1B_1)	$\frac{A_1 \times B_1}{N} = \frac{42 \times 60}{120}$	21	12
(A_2B_1)	$\frac{A_2 \times B_1}{N} = \frac{30 \times 60}{120}$	15	18
(A_3B_1)	$\frac{A_3 \times B_1}{N} = \frac{48 \times 60}{120}$	24	30
(A_1B_2)	$\frac{A_1 \times B_2}{N} = \frac{42 \times 60}{120}$	21	30
(A_2B_2)	$\frac{A_2 \times B_2}{N} = \frac{30 \times 60}{120}$	15	12
(A_3B_2)	$\frac{A_3 \times B_2}{N} = \frac{48 \times 60}{120}$	24	18

f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
12	21	−9	81	3·9
18	15	3	9	0·6
30	24	6	36	1·5
30	21	9	81	3·9
12	15	−3	9	0·6
18	24	−6	36	1·5
				$\chi^2 = 12$

Given, χ^2 table value at 5% level of significance

at $df_2 = 5.99$

$\therefore \boxed{\chi^2_{cal} > \chi^2_{tab}}$ i.e. $(12 > 5.99)$

Since, the difference is significant,

$\therefore$ Hypothesis is rejected.

---

6 In case of sample investigation there are two types of errors :

1. Sampling errors
2. Non-sampling errors.

Sampling error :- The sampling errors are inherent and unavoidable. They occur due to the following reasons –

1. Only a small part of the population is studied and hence the results differ from the census result.
2. Different samples taken from same population but the condition may varry.

There are two types of sampling errors:

1. Biased Sampling errors: They are also called Systematic errors, constant errors or cummulative or persistant errors as they tend to be in the the same direction and increase when the number of item increases. The main reasons for biased errors are:
   a) The method of collection is defective
   b) The arrangement of data is inappropriate
   c) The analysis of data is not proper.

2. Unbiased Sampling errors: These errors are also called compensatory, accidental or random errors. Such errors arise in the normal cause of investigation on account of chance.
They occur accidently without any bias or prejudice. At times they are compensating in nature and thus leave little effect on the general results.

Sampling errors can be reduced by adopting a suitable sampling procedure.

Non-sampling error :-

The errors that arise in a sample investigation, not because of sampling but also appear in a census investigation are known as non-sampling errors. The non-sampling errors occurs due to defective frame, faulty selection of sample units, wrong use of sampling techniques etc.

Such error may occur in any stage of investigation — defective defination of population, incomplete questionnaire, inappropriate data collection, presentation etc.

These errors increase with the increase in the number of items and it includes the biases and mistakes.

Non-sampling errors may be controlled or reduced by careful designing of the questionnaire, intensive trainning of the enumerators, having better supervision, supplying better equipments etc.

---

7. Need for a seperate analysis for test of significance in small samples :

When the number of items in a sample is less than 30, we consider it as small sample. Small samples are required generally when the data is obtained in laboratories or by experiments which requires huge expenditure and consume more time, and when drawing large samples is not possible.

The basis of analysis of large sample is not applicable in case of small samples and seperate analysis is needed for the following reasons : • In small samples, the sample variance is not a reliable or unbiased estimate of population standard deviation as in large samples.

• Small samples do not confirm to the laws of ~~inert~~ inertia of large numbers.

• The interval for population mean given by $\bar{x} \pm 1.96\sigma_{\bar{x}}$ is not true in case of small samples.

Types of test in small samples

There are three main tests of significance for small samples:

(1.) t-test based on t-distribution

(a) $t = \frac{\bar{x} - \mu}{S}$ where,

$\bar{x}$ = Sample mean

μ = population mean

S = Standard deviation of sample

Thus, computed value of t is compared with table value of t to test the significance for judgement of hypothesis.

(b) $$t = \frac{\bar{x}_1 - \bar{x}_2}{S}\sqrt{\frac{N_1 * N_2}{N_1 + N_2}}$$

$$S = \sqrt{\frac{\Sigma d_1^2 + \Sigma d_2^2}{n_1 + n_2 - 2}}$$

The above test is for significance of the difference between two sample means.

(c) $$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

The above test is for significance of coefficient of correlation in small samples.

(2) Variance Ratio Test : F-Test

$$F = \frac{S_1^2}{S_2^2} \; ; \; S_1^2 = \frac{\Sigma (x_1 - \bar{x}_1)^2}{n_1 - 1}, \; S_2^2 = \frac{\Sigma (x_2 - \bar{x}_2)^2}{n_2 - 1}$$

where, S_1^2 is used for the larger variance in the two samples.

(3) Fisher's Z-Test

(a) To test the significance of correlation coefficient. (difference between observed value and assumed value of r)

$$Z_r = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right)$$

$$Z_p = 1.1513 \log_{10} \left(\frac{1+p}{1-p} \right)$$

$$SE_z = \frac{1}{\sqrt{n-3}}$$

$$W = \frac{Z_r - Z_p}{SE_Z}$$

(b) Test of significance between two sample coefficients of correlation

$$SE_{Z_1 - Z_2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

$$W = \frac{Z_1 - Z_2}{SE}$$

8

Analysis of variance is a statistical technique, with help of which total variation, is partitioned into variation caused by each set of independent factors and homogeneity of several means is tested.

$$\text{Variance Ratio, } F = \frac{\text{Variance between Samples}}{\text{Variance within Samples}}$$

i.e.. $$F = \frac{\text{Mean Sum of Squares of deviation between Samples}}{\text{Mean sum of squares of deviation within Samples.}}$$

Significance of F-ratio in analysis of variance :

The calculated value of F is compared with the tabulated value of F.

If $F_c > F_t$, i.e. if calculated value of F exceeds the tabulated value of F, it states that the difference among sample means is significant and conclude that all the population means are not equal.

If $F_c < F_t$, the difference among the sample means is not significant, thus accept the hypothesis that means of the populations are equal.

Prepared by :-

Sumona Bhattacharya
Dept. of Commerce and Financial Studies
Bilaspur University, Bilaspur (C.G.)